Automatic identification of physical activity intensity and modality from the fusion of accelerometry and heart rate data

Journal:	*Methods of Information in Medicine*
Manuscript ID	ME15-01-0130.R1
Manuscript Type:	Original Article
Date Submitted by the Author:	26-Mar-2016
Complete List of Authors:	García-García, Fernando; Universidad Politécnica de Madrid, Bioengineering and Telemedicine Group; Networking Research Center on Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN) Benito, Pedro; Universidad Politécnica de Madrid, Laboratory of Exercise Physiology, Facultad de Ciencias de la Actividad Física y del Deporte (INEF) Hernando, Elena; Universidad Politécnica de Madrid, Bioengineering and Telemedicine Group; Networking Research Center on Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN)
Keywords:	physical activity intensity, exercise modality, accelerometer, heart rate, clustering

Title:

Automatic identification of physical activity intensity and modality from the fusion of accelerometry and heart rate data

Running title:

Physical activity identification, accelerometry & HR

Authors:

F. García-García

1. Bioengineering and Telemedicine Group, Universidad Politécnica de Madrid; Spain
2. Networking Research Center on Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN); Madrid, Spain

P. J. Benito

3. [On behalf of the PRONAF Study Group]. Laboratory of Exercise Physiology, Facultad de Ciencias de la Actividad Física y del Deporte (INEF), Universidad Politécnica de Madrid; Spain

M. E. Hernando

1. Bioengineering and Telemedicine Group, Universidad Politécnica de Madrid; Spain
2. Networking Research Center on Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN); Madrid, Spain

Corresponding author:

Fernando García-García
e-mail: fgarcia@gbt.tfo.upm.es
Address: ETSI Telecomunicación (B-303), Avda. Complutense 30. 28040, Madrid (Spain)
Telephone: +34 915495700 ext. 3407 Fax: +34 913366828

Summary:

Background: Physical activity (PA) is essential to prevent and to treat a variety of chronic diseases. The automated detection and quantification of PA over time empowers lifestyle interventions, facilitating reliable exercise tracking and data-driven counseling.

Methods: We propose and compare various combinations of machine learning (ML) schemes for the automatic classification of PA from multi-modal data, simultaneously captured by a biaxial accelerometer and a heart rate (HR) monitor. Intensity levels (low/moderate/vigorous) were recognized, as well as for vigorous exercise, its modality (sustained aerobic/resistance/mixed). In total, 178.63 h of data about PA intensity (65.55% low/18.96% moderate/15.49% vigorous) and 17.00 h about modality were collected in two experiments: one in free-living conditions, another in a fitness center under controlled protocols. The structure used for automatic classification comprised: a) definition of 42 time-domain signal features, b) dimensionality reduction, c) data clustering, and d) temporal filtering to exploit time redundancy by means of a Hidden Markov Model (HMM). Four dimensionality reduction techniques and four clustering algorithms were studied. In order to cope with class imbalance in the dataset, a custom performance metric was defined to aggregate recognition accuracy, precision and recall.

Results: The best scheme, which comprised a projection through Linear Discriminant Analysis (LDA) and *k*-means clustering, was evaluated in leave-one-subject-out cross-validation; notably outperforming the standard industry procedures for PA intensity classification: score 84.65%, versus up to 63.60%. Errors tended to be brief and to appear around transients.

Conclusions: The application of ML techniques for pattern identification and temporal filtering allowed to merge accelerometry and HR data in a solid manner, and achieved markedly better recognition performances than the standard methods for PA intensity estimation.

Keywords:

Physical activity intensity, exercise modality, accelerometer, heart rate, clustering.

----

1. Introduction

1.1. Motivation

The promotion of physically active lifestyles is essential for the prevention and treatment of a wide variety of chronic diseases with large prevalence and impact on public health: obesity [1], type 2 diabetes [2], cardiovascular [3] and respiratory diseases [4], hypertension, cancer [5] and depression, among others. Hence, a reliable automated detection and quantification of physical activity (PA) –along with its evolution over time– can empower lifestyle interventions on populations by facilitating: a) the logging of exercise data and its subsequent analysis, b) the monitoring of patients' degree of adherence and compliance with prescribed PA plans, and/or c) the provision of individually tailored, data-driven feedback to patients and caregivers.

In the particular application scenario of type 1 diabetes, distinguishing not only PA intensity but also its predominant exercise modality, may be of major interest since different PA modalities induce remarkably distinct acute responses in glycaemia [6]. As a consequence, the adjustment of therapeutic strategies for insulin infusion can be enhanced by the provision of objective, quantitative information about PA intensity and modality [7][8].

Gold standard techniques to measure PA (i.e. doubly labelled water $^2H_2{}^{18}O$ and indirect calorimetry, via the analysis of pulmonary gas exchanges), despite their high reliability and validity, are only appropriate in laboratory scenarios. Practical issues include among others: cost inefficiency, obstructiveness of the equipment –calorimeters, in particular– and lack of temporal resolution, in the case of doubly labelled water [9].

In field use, accelerometers and heart rate (HR) monitors are reasonable, practical alternatives; each technique with its respective strengths and limitations. Accelerometers record displacements in one or multiple axes with notable sensitivity, although their resulting overall recognition accuracy depends substantially on whether the particular activity under monitoring involves or not movement of the body part to which the accelerometer is attached [10]. On the other hand, HR monitors track a physiological response to exercise which exhibits a fairly close relationship with oxygen consumption rates and energy expenditures (EE); this relation being approximately linear at moderate intensity ranges: around 110–150 beats/min [10]. Nevertheless, the cardiovascular response to high-volume resistance training differs notably from such behavior [11]. Other inter- and intra-subject factors may play a role on HR responses: e.g. age, sex, degree of fitness, stress or medication, among others. In this regard, a number of researchers from the field of sports sciences advocate for the fusion of accelerometry and HR measurements, as complementary sources of information describing both mechanical and physiological aspects of PA patterns [10][12][13][14].

1.2. Related works

Machine learning (ML) techniques have been extensively and successfully applied in literature ~~to~~ for the recognition of specific activities and/or body postures. The ~~vast~~ majority of works following this approach selected a closed, fixed set of target activity options (commonly: lie, sit, stand, walk, run and/or bicycle) and discerned among them based on accelerometry measurements only. Closely related~~,~~ ML-oriented research topics include accelerometry-enabled gait analysis [15][16] and fall detection [17]; where the targeted population is mainly elderly subjects who, given their frailty status, may not benefit from PA monitoring under more general scenarios.

~~A~~ From an algorithmic point of view, a wide variety of ML classification ~~algorithms~~ schemes have been explored, including: *k*-nearest neighbors, Naïve Bayes (NB) classifiers and C4.5 decision trees [18]~~[15]~~, logistic regression [17][19], Multi-Layer Perceptron (MLP) neural networks ~~[16]~~[20][21]~~[21][21][17]~~, Support Vector Machines (SVM) [22]~~[22][22][18]~~[23]~~[23][23][19]~~, Random Forests [24]~~[24][24][20]~~ or Boosting ensembles [25]~~[25][25][21]~~, among others. Fewer authors incorporated HR data, coming up with remarkably dissimilar results: study [26]~~[26][26][22]~~ discarded the HR signal for not increasing sufficiently the accuracy obtained by their C4.5 and NB classifiers; whereas notorious performances in activity-specific recognition in laboratory environments were reported by works [27]~~[27][27][23]~~[28]~~[28][28][24]~~.

Other ~~group~~ set of ~~works~~ publications opted for deploying activity-specific recognition systems as a preliminary stage for their subsequent ~~energy expenditure~~EE estimators, the latter specifically tuned for each recognized activity [29]~~[29][29][25]~~[30]~~[30][30][26]~~. Studies [31]~~[31][31][27]~~[32]~~[32][32][28]~~ reduced their ~~sets~~ ranges of PA options to four and five activity conglomerates, respectively; whereas [33]~~[33][33][29]~~ incorporated ECG information and grouped activities by intensity ~~ranges~~levels.

A third family of approaches consist in the activity-independent ranking of PA intensities. ML literature in this regard is scarcer: works [34]~~[34][34][30]~~[35]~~[35][35][31]~~ developed MLP-based energy estimators operating on accelerometry data and then assigned activities to a certain intensity level in accordance; whereas [28]~~[28][28][24]~~ adapted their activity-specific recognition schemes to serve as an activity-independent PA classifier.

~~2.~~1.3. Objectives

Formatted: Indent: Left: 0", First line: 0"

In this work we propose and compare a series of combinations of ML algorithms explicitly designed to recognize PA patterns along time, identifying not only their intensity level, but also their predominant exercise modality ~~(~~: an aspect which, to the best of our knowledge, has not yet been addressed in literature~~)~~. To do so, multi-modal data from accelerometry and HR will be simultaneously merged.

~~3.~~2. Materials

~~3.1.~~2.1. Equipment

ActiTrainer (ActiGraph, USA) accelerometry devices were ~~selected~~ used here for data collection. – The selection of ActiTrainer was based on two main criteria~~,~~ which stand out as distinctive features of this equipment, namely: a) its ~~due to their~~ thoroughly documented validity and field reliability [36]~~[36][36][32]~~[37]~~[37][37][33]~~[38]~~[38][38][34]~~; as well as for b) ~~their~~ its capability to establish wireless communication with Polar Wearlink (Polar Electro, Finland) HR monitors in a transparent manner, thus allowing for a straightforward~~,~~ and fully synchronous capture of accelerometry ~~and~~plus HR signals without the need for ~~performing~~ subsequent synchronization procedures. Additionally, ActiGraph's pedometer functionality [39]~~[39][39][35]~~ was also exploited.

As other comparable commercial accelerometry devices dedicated for PA monitoring in sports sciences [40]~~[40][40][36]~~, ActiTrainer produces its outcome measurements in the form of 'activity counts' ~~through~~ via a proprietary algorithm~~,~~ which integrates the rectified raw accelerometry signal over an epoch with configurable duration [41]~~[41][41][37]~~.

During our experiments, participants wore the ActiTrainer biaxial accelerometer (86×33×15 mm size, 51 g weight, ±3G dynamic range, 30 Hz sampling frequency) tightly attached to their waist, with main and secondary axes respectively oriented in vertical and antero-posterior directions, as ~~suggested~~ recommended by manufacturer's guidelines [41]~~[41][41][37]~~. For the HR signals to be handled appropriately, ActiTrainer's firmware imposed epochs with a minimum duration of 10 s, which was the choice here for the sake of maximal temporal resolution.

~~3.2.~~2.2. Experiments and participants

Two distinct data collection experiments were conducted: one covered free-living conditions in ambulatory scenarios, whereas the other was carried out in a controlled laboratory environment.

Experiment A enrolled seven healthy individuals: 3 males and 4 females, with an age range of 25–43 years old and diverse lifestyles: varying from sedentary to regular sport practice. Volunteers received instructions on how to wear the sensors and were requested to produce detailed written reports describing the timing and intensity of their activities. Subjects were encouraged to record: i) daily life situations, ii) the use of means of transportation, as well as iii) their preferred PAs at self-selected intensities.

Experiment B took place in a fitness center, under strict timing control and direct supervision by the research team. Three alternative types of circuit were considered: two comprised upper- and lower-limb resistance exercise (i.e. strength/weight training; e.g. shoulder press, barbell roll, squat) in either fitness machines or with free weight (Johnson Health Tech Iberica, Spain), whereas the other circuit combined free weight and aerobic exercise –treadmill running– performed in short alternating bouts. Each 64-min session started with a 5-min warm-up phase (mild treadmill or elliptical walk), plus a preliminary circuit lap with duration 7 min 45 s and light load (30% of each subject's 15 repetitions maximum, RM). Thereafter, three more circuit laps were performed with high load (at 70% of 15 RM). Laps were separated by 5 min 'active recovery' periods, i.e. walking. A complete description of exercises, circuits and protocols (PRONAF Study) can be found elsewhere [42]~~[42][42][38]~~.
Nine subjects took part: 6 males and 3 females, with an age range of 20–49 years old. Three were healthy, moderately active males; whereas the remaining six individuals suffered moderate overweight: BMI=28.1±1.3 kg/m^2 (mean±SD). Informed consent was obtained in all cases.

~~3.3.~~2.3. Dataset overview

A heterogeneous set of free-living activities in ambulatory scenarios was acquired through Exp. A. This included assorted daily life situations (e.g. sleep, household and office work) and the use of transportation (bus, car, subway and elevator); along with a notable variety of exercises at different intensities, including: walking at various paces, dancing, jogging, vigorous endurance running, karate, football (soccer) and mountain bike. The elimination of ambiguously annotated periods yielded a total of 148.50 h of usable data, split in 72 sessions (Table 1). Session duration was 119 [93–149] min (median [inter-quartile range]), with minimum at 18 min and maximum at 216 min. The detailed distribution by individuals is depicted in Figure 1 (panel A), with a median time of 10.40 h per participant. Of note, the subject with identification #A7 contributed with merely 18 min; whereas on the contrary, subject #A5 (a moderately active, 26-year-old male) was very enthusiastic and participative in the data collection procedure, contributing with a total of 81.87 h.
In contrast, Exp. B in the controlled laboratory environment was markedly more homogeneous with respect to the exercise activities covered. Differences in the volumes of data for each subject had two sources:

i) The number of sessions in which each volunteer participated, which depended on his/her availability for the experiment: five overweight individuals (IDs #B4 to #B7 and #B9) completed all three circuit versions, on different days and random order; whereas one healthy subject (#B1) exercised for two sessions, and the remaining three participants (two healthy males #B2, #B3, plus one overweight female #B8) completed one circuit; randomly assigned in all cases. Overall, a total of 20 sessions were registered.

ii) The pre- and post-exercise resting time that was recorded in addition to the 64-min circuit protocol.

TABLE 1 HERE

2.4. Dataset labelling

Data intervals were each assigned to different PA intensity level groups attending to their Metabolic Equivalent of Task (MET) values. The concept of METs was particularly useful ~~to us~~here, since it allows to quantify PA energy costs as a multiple of each subject's basal metabolic rate (BMR) [43], which is feasible~~. This occurs~~ because body mass tends to influence~~s~~ both total PA-induced ~~energy expenditures~~EE and BMR in a comparable manner. In particular, 1.0 MET is by convention made equivalent to 1.0 kcal/min·kg^{-1} [44].
For the purpose of this work, we ~~manually~~ separated data items into three PA intensity classes, using standard range criteria by the American College of Sports Medicine [45] ~~for the intensity ranges~~~~according to Ainsworth's Compendium [39]. Standard intensity ranges [40] were employed~~, namely: i) low-intensity activities, below 3 MET ~~Metabolic Equivalents of Task (MET)~~ –which therefore include~~d~~ sedentary situations–, ii) moderate PA (i.e. 3–6 MET), and iii) vigorous exercise

(>6 MET). Ground truth METs were assigned ~~here~~ to match those values provided by Ainsworth's Compendium [46], and on the basis of participants' reference annotations ~~(~~or researchers' laboratory notes ~~notes,~~ in ~~the specific case of~~ Exp. B~~)~~.

~~In addition, 17.00 h of~~For the specific case of vigorous ~~PA~~exercise ~~data (Table 1, lower part) incorporated further information about the main exercise modality involved: either sustained aerobic, mixed or resistance activity~~, a further distinction was considered in terms of PA modality, again with three possible classes: aerobic, mixed or anaerobic. This approach was conceived to ~~allow us to~~ ascertain which metabolic pathway (aerobic vs. anaerobic) is predominant at each moment in vigorous PA, or whenever both mechanisms had ~~notable~~comparable contributions (mixed class). The rationale for this discernment resides in the fact that the proportions of each pathway's contribution~~,~~ –along with PA intensity~~–,~~ play an important role on the actual physiological and metabolic responses to exercise, for example in patients with type 1 diabetes [6]. In the context of Exp. A, ~~which was~~ taking into account that it was carried out under unsupervised~~,~~ free-living conditions; for the purpose of studying the predominant PA modality~~,~~ ~~here~~ we decided to include only prototypical situations: ~~(in this specific case:~~here merely endurance running and bicycle; both allocated into aerobic modality data~~)~~. These data recordings accounted for a total of 6.80 hours spread in 5 sequences, all performed by the same volunteer (ID #A5). Conversely, the laboratory setup of Exp. B allowed for a more precise reference concerning PA modality, therefore being this information ~~therefore~~ available for all 10.20 h of vigorous PA (20 sessions, 9 participants). Of these Exp. 2 data, 3.45 h corresponded to mixed PA modality; whereas the remaining 6.75 h were predominantly anaerobic. ~~It may be worth mentioning that, in the context of sports science,~~Of note, considerable research in sports sciences is currently focused on weight loss training programs ~~combining~~which combine aerobic and strength (i.e. anaerobic) exercise, e.g. [47]. On the other hand, aerobic activity is fairly more common in self-selected leisure sport.

FIGURE 1 HERE

~~4.~~3. Methods

In this work we investigate a pipeline of ML algorithms which, in broad terms, correspond to the steps outlined by review [48]~~[48][48][41]~~ for activity-specific PA recognition. Major novelties with respect to [48]~~[48][48][41]~~ are: i) the use of unsupervised clustering techniques for pattern extraction and vector quantization [49]~~[49][49][42]~~; and ii) a final Hidden Markov Model (HMM), functioning as an extra temporal filter module which exploits time redundancies in the series of data.

~~4.1.~~3.1. Signal processing and feature definition

Biaxial accelerometry, step counts and HR data were simultaneously recorded and processed. Signals were first divided into non-overlapping window segments with duration 2 min (N=12 epochs at 10 s each). In this regard, several alternative window lengths were tested in a preliminary stage, where these 2 min were found to provide a satisfactory trade-off between: a) capturing information-rich patterns in the signals, task for which longer analysis segments would be desirable; and b) sufficient temporal resolution (i.e. low time granularity), which implied short windows.

In addition, activity counts in the main and secondary axes $\boldsymbol{a_1}$, $\boldsymbol{a_2}$ were combined to form an extra magnitude, called here 'pseudonorm' counts $\boldsymbol{a_{pn}}$, and defined as $a_{pn} = \sqrt{a_1^2 + a_2^2}$. For every window and each signal (namely: accelerations $\boldsymbol{a_1}$, $\boldsymbol{a_2}$, $\boldsymbol{a_{pn}}$; ActiTrainer's step counts $\boldsymbol{st}$, and $\boldsymbol{hr}$ i.e. Polar's HR measurements), time-domain statistical descriptors were computed. These features included:

- Mean and standard deviation,
- Mean versus median difference (given that median is less sensitive to eventual outliers),
- Maximum value and total range,
- Signal variability within a window, computed as the accumulation of absolute differences between a sample $\boldsymbol{x(t_i)}$ of signal $\boldsymbol{x}$ at time $\boldsymbol{t_i}$ and its previous value:

$$\sum_{i=2}^{N} |x(t_i) - x(t_{i-1})|$$

- A coefficient of dispersion with respect to the median value $\boldsymbol{x_{med}}$ within the window ($\boldsymbol{x_{med}} \neq 0$):

$$\frac{1}{N}\sum_{i=1}^{N} \frac{|x(t_i) - x_{med}|}{x_{med}}$$

We observed that features computed from acceleration counts (i.e. those derived from signal $\boldsymbol{a_1}$, $\boldsymbol{a_2}$ and $\boldsymbol{a_{pn}}$; but neither from ***st*** nor ***hr***) yielded values which spanned across several orders of magnitude. This issue could pose a challenge for the subsequent ML schemes, especially those with learning based on distances. To face it, a truncated signed logarithmic transformation was applied [50]~~[50]~~[50]~~[43]~~:

$$f(x) = \begin{cases} sign(x)\log|x|, & |x| \geq 1 \\ 0, & otherwise \end{cases}$$

The sign considerations were necessary to accommodate negative feature values, which given their definition (and taking into account that $\boldsymbol{a_1}$, $\boldsymbol{a_2}$ are by definition non-negative integers), could only arise in features based on mean-vs-median differences.

Additional ad hoc features were also used, namely:

- Three paired products between mean ‘pseudonorm’ counts $\boldsymbol{a_{pn}}$ (log-transformed), mean HR and mean number of steps; along with the combined triple product of these magnitudes
- Histogram of counts in the main axis $\boldsymbol{a_1}$, where ActiGraph’s default cutpoints [51]~~[51]~~[51]~~[44]~~ were selected as the reference for setting the extremes of the binning intervals: More precisely, cutpoints were proportionally rescaled in order to adapt them to our 10 s epochs, instead of the original 60 s [51]~~[51]~~[51]~~[44]~~.

In total, 42 time-domain features were computed.

~~4.2.~~3.2. Dimensionality reduction

This work explores various techniques to retain as much information as possible from the original high-dimensional feature space, while reducing the number of variables and hence, the complexity of the automated learning task. In this manner, unnecessary redundancy (as well as, to some extent, also noise) is cancelled out from the dataset.

Prior to the dimensionality reduction stage itself, and given that our features were clearly not commensurate, the feature space was standardized: subtracting sample mean in each dimension and normalizing by its standard deviation. This also prevented relevant information contained in low-scaled features from being shaded by variables with high variance.

~~4.2.1.~~3.2.1. Reduction by projection

New features were generated through linear combinations of the original ones:

i) Principal Component Analysis (PCA) projects data onto a feature subspace preserving as much randomness as possible from the original high-dimensional dataset. Such subspace is defined by the eigenvectors with largest eigenvalues from the sample covariance matrix [52]~~[52]~~[52]~~[45]~~; although given our pre-standardization step to cope with non-commensurability, this covariance matrix coincided with the correlation matrix of the untransformed features [52]~~[52]~~[52]~~[45]~~.

ii) Linear Discriminant Analysis (LDA) is optimized to transform the data space to enhance class separability and discriminatory information once projected. Consequently, LDA demands ground truth class assignments; in contrast to PCA, which is an unsupervised scheme. The core of LDA relies on solving a generalized eigenvalue problem where the within- and between-class scatter matrices serve as indicators of class separability [52]~~[52]~~[52]~~[45]~~. The dimensionality of the output subspace equals the number of classes under consideration minus one.

~~4.2.2.~~3.2.2. Reduction by feature selection

Instead of computing appropriately weighted linear combinations of features, as in PCA or LDA (section 4.2.1); feature selection methodologies search for a suitable subset of features in accordance to certain optimality criterion [53]~~[53]~~[53]~~[46]~~[54]~~[54]~~[54]~~[47]~~.
In this work we adopted the so-called filter approach [54]~~[54]~~[54]~~[47]~~. Filters function by selecting features according to their ability to maximize a target merit function, in turn defined to quantify the appropriateness of a candidate subset based on criteria stemming from information theory. Here we employed the so-called Minimum Redundancy-Maximum Relevance (mRMR) criterion [55]~~[55]~~[55]~~[48]~~[56]~~[56]~~[56]~~[49]~~. At each iteration, one feature is selected in a greedy, near-optimal manner [55]~~[55]~~[55]~~[48]~~, to optimize the trade-off between: a) additional information supplied by such feature, thus maximizing relevance; and b) redundancy with respect to the other variables already filtered in, hence minimizing redundancy. We implemented the algorithms as described by work [56]~~[56]~~[56]~~[49]~~ in both their continuous and discrete/categorical versions; using respectively the FCQ (F-test Correlation Quotient) and MIQ (Mutual Information Quotient) metrics to rank candidate features. Given that our variables were continuous instead of categorical, for the discrete version of mRMR we binned each variable into eight intervals, symmetrically distributed around the sample mean and with widths equal to one half of the sample standard deviation. In the standardized feature space, this definition of binning intervals was equivalent to the following ranges: $(-\infty,-3/2)$; $[-3/2,-1)$; $[-1,-1/2)$; $[-1/2,0)$; $[0,1/2)$; $[1/2,1)$; $[1,3/2)$ and $[3/2,+\infty)$.

The PCA projection was capable of covering >90% of total variance (exactly 91.43%) with a subspace built on only 8 eigenvectors; whereas the LDA projection generated two discriminants: i.e. $C-1$, with $C=3$ being the number of PA classes under consideration here. For the two mRMR filter-based feature selection procedures, for the sake of comparability with respect to PCA, the target number of features was set fixed to 8. In particular, the subsets selected by both the continuous and categorical/discrete versions of mRMR coincided moderately, with 5 out of 8 features in common: two features based solely on HR information (mean and maximum HR), one in relation to the pedometer functionality (mean step count), another feature related to the histogram of counts and the combined product of mean HR and mean 'pseudonorm' counts $\boldsymbol{a}_{pn}$ (log-transformed).

~~4.3.~~3.3. Data clustering

We explored four unsupervised clustering techniques to automatically discover relevant underlying patterns in the multi-modal signals. Data were grouped into clusters which, in principle, lack intuitive interpretation solely by themselves; but constituted the input for the subsequent temporal HMM filtering stage. The four algorithms under study here were [49]~~[49]~~[49]~~[42]~~:

i) Classical *k*-means clustering.

ii) Gaussian Mixture Models (GMM), which approximate the probability density function of data as a mixture –i.e. weighted linear combination– of multidimensional Gaussian distributions whose parameters (means, covariance matrices and weights) were estimated via an Expectation-Maximization algorithm. Each cluster corresponded to a Gaussian component.

iii) Agglomerative hierarchical clustering, with Euclidean distance criterion and Ward's linkage. These particular distance and linkage functions were chosen during a preliminary tuning stage as the options providing highest performance.

iv) Self-Organizing Maps (SOM), subsequently followed by an agglomerative hierarchical clustering on the SOM neurons; instead of on the whole point cloud (as it was done in the previous scheme), with a dramatic decrease in terms of computational load. The topology of the SOM neural network (hexagonal grid), the number of neurons (20×20 here), and the linkage criterion for hierarchical clustering (average) were also tuned preliminarily.

~~4.4.~~3.4. Temporal filtering

HMMs have been employed in several activity-specific PA classifiers in literature, mainly to characterize transitions between body postures [25]~~[25]~~[25]~~[21]~~[57]~~[57]~~[57]~~[50]~~[58]~~[58]~~[58]~~[51]~~[59]~~[59]~~[59]~~[52]~~. Conversely, in our approach we use a HMM to exploit the strong time interdependency between neighbor windows which exists in our scenario, gaining benefit of this temporal redundancy for the sake of robustness in classification. Indeed, PA intensity and modality during ambulatory exertions did not often tend to change in a quickly fluctuating manner; on the contrary, variations were most often gradual, with long-term trends maintained stable for periods that in general, broadly exceeded the duration of our 2 min analysis window.

Using HMM terminology, for our purpose we assumed the observable Markovian process to correspond to cluster assignments; whereas the hidden process of interest would be the PA classes which most likely generated the observed sequence of cluster assignments. Therefore, once the HMM had been trained (i.e. its emission and transition matrices estimated), the reconstruction of the most likely temporal succession of hidden states –PA classes– was achieved by applying Viterbi's algorithm [60]~~[60]~~[60]~~[53]~~.

~~4.5.~~3.5. Other practical aspects

~~4.5.1.~~3.5.1. Performance evaluation and model selection

To assess the performance attained by the algorithms, and in order to select the optimal model: i.e. the most appropriate combination of dimensionality reduction method and clustering scheme, we declined to employ overall classification accuracy as our single performance metric. The rationale is that accuracy cannot fairly reflect by itself achievements for under-represented classes [61]~~[61]~~[61]~~[54]~~, an issue which is critical in our dataset due to the marked imbalance of samples representing each PA class (Table 1). Instead, we defined a custom compound performance score averaging: a) classification accuracy, and b) *f*-Measures for each class under consideration.

$$Score = \frac{1}{C+1}\left[Accuracy + \sum_{c=1}^{C} fMeasure(Class_c)\right]$$

where *C* represents the total number of classes, and *f*-Measure is a performance metric commonly used in information retrieval, defined as the harmonic mean between precision (i.e. positive predictive value) and recall (i.e. sensitivity):

$$fMeasure = 2\frac{Precision \cdot Recall}{Precision + Recall}$$

$$Precision = \frac{TruePos}{TruePos + FalsePos}$$

$$Recall = \frac{TruePos}{TruePos + FalseNeg}$$

For reliability in the assessments of performance, a stratified 10-fold cross-validation (CV) was carried out for the task of PA intensity classification; repeating the whole process n=30 times in order to ascertain its variability. When allocating data across CV folds, stratification was not done on a sample basis, because that procedure would have dismissed any information about temporal trends underlying in consecutive windows. Instead, we implemented a block-based or sequence-based stratified CV (SBSCV): each sequence –i.e. each block of data recorded during the same PA session– was randomly assigned to a certain fold. Subsequently, folds were checked to verify if: a) the relative occurrence of the *C* classes in each fold did not notably deviate from the overall class priors (Table 1); and if b) the total number of data samples were similar for all folds. In the event that any of these two conditions was violated, the randomization was repeated until both conditions were satisfied.

For the study of performances in the classification of PA modality, we conduced 5-fold SBSCV instead of 10-fold, due to the restrained number of PA sequences available concerning modality.

~~4.5.2.~~3.5.2. Parameter tuning

For each of the learning algorithms in section 4.3, an optimal number of clusters needed to be established before the training phase could take place. For this tuning procedure, the space of possible parameter values was explored using a grid search strategy, with nested 10×10-fold SBSCV conducted on the PA intensity data, aiming to maximize the target custom score.

~~5.~~4. Results

~~5.1.~~4.1. Model selection

Table 2 summarizes the outcome performance metrics for PA intensity classification during the 10-fold SBSCV model selection procedure, as obtained by all of the combinations between dimensionality reduction schemes and clustering algorithm. The highest performance was achieved by the LDA+SOM+HMM scheme, with a score of 84.67±0.57% and an accuracy of 89.09±0.42% (mean±SD); followed by LDA+*k*-means+HMM and LDA+hierarchical clustering+HMM, with slightly lower –although virtually identical– average scores. For a definitive model selection, these three options were trained for the recognition of PA modality during vigorous exercise, yielding respective scores of 77.94±5.93%, 82.61±7.16% and 80.08±4.90%. Consequently, we selected the scheme comprising LDA+*k*-means+HMM, whose detailed SBSCV performance in both PA intensity and modality classification is shown in Table 3.

TABLE 2 HERE

TABLE 3 HERE

~~5.2.~~4.2. Final performance evaluation

For a definitive assessment of the classification performance achievable by the selected LDA+*k*-means+HMM scheme, we carried out a leave-one-~~subject~~individual-out CV (~~LOSOCV~~LOIOCV), in which data sequences belonging to all individuals except one were employed in successive turns to train the algorithm, whereas the remaining data from the unseen subject served for evaluation purposes.

For the PA intensity classification task (Table 4), a total score of 84.65% was obtained, with accuracy 88.86% and respective *f*-Measures 95.59%, 72.28% and 81.86%. Of note, when subject #A5 (who contributed with 45.83% of the total volume of data) was left aside from the training set and used for testing, the system reached outstanding recognition: score 96.69%. However, we could not perform ~~LOSOCV~~LOIOCV for PA modality recognition, since only this individual supplied data regarding the sustained aerobic modality.

TABLE 4 HERE

FIGURE 2 HERE

Figure 2 depicts several example sequences resulting from the ~~LOSOCV~~LOIOCV classification of PA intensities, plotted versus ground truth. In general, mismatches tended to appear in either two situations: i) around transitions in the ground truth (e.g. at approximately 1½, 2½, 3, 6, 14 or 17 h in panel A; or at around 3½, 12 or 17 h in panel B); or ii) where ground truth fluctuated rapidly in the short-term (panels C, D). These latter fluctuations had their origin in the specific exercise scheduling for Exp. B (PRONAF Study protocol); where three vigorous bouts of 7¾ min duration were alternated with 5 min 'active recovery' periods. However, long-term discrepancies rarely occurred, as errors tended to be brief in duration.

~~5.3.~~4.3. Comparison with other methods

~~5.3.1.~~4.3.1. Standard approaches in the application domain

The most widespread technique to identify PA intensity levels is by means of 'cutpoints' (i.e. thresholds), applied on the activity counts recorded for the main axis a1. This simple solution is the industry de facto standard for ActiGraph devices [41]~~[41]~~[41]~~[37]~~ and other commercial accelerometry equipment, extensively used by sports scientists.

For the sake of comparison, we applied this 'cutpoints'-based methodology on the dataset under study here, accumulating the activity counts for axis a_1 over 1 min epochs, in order to match the original proposal [51]~~[51]~~[51]~~[44]~~. This yielded a modest performance score of 63.60%, with accuracy 77.45% and *f*-Measures 87.19%, 61.01% and 28.56% respectively for low, moderate and vigorous PA intensity categories. We also examined a widely accepted MET estimation formula, suitable for ActiGraph accelerometers [62]~~[62]~~[62]~~[55]~~, which is in turn based on non-linear regressions; and grouped outcomes into their corresponding PA intensity level. Results in this case were slightly inferior to the former: score 63.15%, accuracy 77.56% and *f*-Measures 87.46%, 62.20% 410 and 25.39%.

Both methodologies presented an important portion of errors when facing vigorous resistance exercise from Exp. B which, being static with respect to waist movements (i.e. the location where the accelerometry sensor was placed), were classified as low intensity.

~~5.3.2.~~4.3.2. *ML baseline comparison*

None of the two techniques introduced and evaluated in the previous section were designed to incorporate HR-related information, an issue which may explain –at least partially– their poor overall performance; and in particular, their failure to detect resistance exercise. In order to compare our main results with schemes capable of using the merge of data from accelerometry and HR, we implemented two basic ML classifiers, specifically: CART decision trees and NBs. These algorithms were applied on a two-dimensional feature space combining: a) the total activity counts in the main axis $\boldsymbol{a_1}$, accumulated over the 2 min window periods, and b) the average HR. These two magnitudes should be, at least intuitively [63]~~[63][63][56]~~, the most informative signal descriptors, covering both mechanical and physiological phenomena in relation to PA.

CART decision trees performed slightly better with activity counts $\boldsymbol{a_1}$ expressed in natural units: score 72.27%, accuracy 80.31% and *f*-Measures 92.01%, 52.94% and 63.82% in ~~LOSOCV~~LOIOCV; whereas NBs worked best when operating on log-transformed $\boldsymbol{a_1}$ counts: score 76.79%, accuracy 82.941% and *f*-Measures 93.08%, 60.63% and 70.49% in ~~LOSOCV~~LOIOCV.

Whereas these two ML methods incorporating HR improved notably the recognition of vigorous resistance patterns from Exp. B with respect to the 'cutpoints'-based industry standard approaches, their overall performance scores were yet markedly inferior to our core proposal.

~~6.~~5. Discussion

In this work we proposed, evaluated and compared a series of combinations of ML algorithms applied to the activity-independent classification of PA intensities, as well as of exercise modality for vigorous periods. These systems worked on data fusing biaxial accelerometry and HR measurements, in order to combine simultaneously motion and physiological aspects of PA, and to overcome the limitations of each measurement technique when used by separate.

The vast majority of ML-based PA classifiers in literature addressed activity-specific recognition problems, having proved their capability to discern with notable accuracy from a closed set of activity options, generally performed in controlled laboratory environments. However, the range of activities embraced by those activity-specific approaches (typically lie, sit, stand, walk, run and/or bicycle) may in practice be excessively limited as to apply these methodologies in the monitoring of free-living ambulatory scenarios, where considerably more heterogeneous PA and non-PA patterns occur. In this regard, our dataset encompassed an appreciable variety of situations, including daily life activities and the use of means of transportation –allowing the system to learn eventual artefacts caused by them–, as well as assorted PAs (e.g.: dancing, karate, soccer or resistance/weight training) which correspond to real-life situations seldom covered by other works.

To address this heterogeneity, an activity-independent classification was adopted, discerning PA intensity levels. Furthermore, this work constitutes~~,~~ –to our knowledge~~,~~ –the first attempt to identify PA modality, an aspect which may be of interest in various applications, e.g. for the management of type 1 diabetes.

The aggregation of accelerometry measurements –reflecting displacement– and HR –as a physiologically sound marker~~,~~ in close relation to PA-induced load on the cardiovascular system– constitutes a powerful strategy for which sports scientists have been advocating [10][13][14]. However, this approach has not gained extensive benefit from the application of ML techniques, since research efforts have been mainly focused on mining exclusively accelerometry data for activity-specific discernments. In this regard, we found counter-intuitive that authors in [26]~~[26][26][22]~~ discarded HR-based features for not improving sufficiently their recognition;

whereas, on the contrary, notorious success in PA intensity classification was reported by work [33]~~[33][33][29]~~ when fusing accelerometry and ECG data. Authors reported outstanding classification rates for a set of activities performed in a laboratory environment; although at the expense of a complex network of sensors (three triaxial accelerometers plus an ECG recorder). In this regard and in order to avoid sensor validity issues during field use, we opted for ActiGraph commercial equipment, which is widely and thoroughly validated by sports scientists and epidemiologists for the ambulatory monitoring of PA in heterogeneous populations. Nevertheless, the 10-s epoch limitation (imposed by ActiTrainer ActiGraph's firmware when registering HR data)~~,~~ prevented ~~us from carrying out~~ any spectral or wavelet analysis on the accelerometry signals, which could however have provided information-rich features [18] ~~[15]~~.

There exist notorious constraints when using 'counts' as source of accelerometry information. First, given that the procedure to generate counts consists in integrating the raw signal along the duration of the epoch, details about waveforms get unavoidably discarded as a consequence of the integration operation. Besides, it is well established that manufacturer-specific design choices for internal pre-processing stages (e.g. signal amplification gains) impede the direct quantitative comparison of 'counts' outcomes across brands; and hinder the overall interpretability of counts from both physical and physiological perspectives [40]. Nevertheless, counts remain to be the most accepted form of output measurement in PA-specific accelerometry devices. On the other hand, general-purpose accelerometers (which could eventually enable us to bypass the limitations associated to counts) lack large field validation studies supporting their use for PA monitoring, regrettably.

We collected a total of 178.63 h of multi-modal data (5359 windows, each with duration 2 min) divided in 92 sequences/sessions from a diverse population formed by 16 subjects (9 males and 7 females, age range 20–49), healthy and overweight individuals with different lifestyles~~–~~. The collection of data resulted notably imbalanced in terms of the number of instances available that corresponded to each class, in particular for the case of PA intensity discernments (moderate and vigorous levels), although not so critically for PA modality. For this reason, we actively enforced stratification during the CV training and evaluation procedures, thus allowing the ML system to learn always from a subset of data in which the proportion of classes is maintained stable across folds, and equivalent to the class priors. Besides, imbalance also lead us to propose a custom score metric which rewards explicitly achievements in underrepresented classes via the summation of *f*-Measures in these classes, and not only accounting for total accuracy. Furthermore, this custom target score drove all of the parameter tuning and model selection processes, hence proactively promoting a specific focus on recognizing minority classes. In this sense, we consider extraordinarily meritorious for the LDA+*k*-means+HMM learning scheme to achieve *f*-Measure outcomes of 71.90% and 82.01% (SBSCV) and 72.28% and 81.86% (LOIOCV) for moderate and vigorous exercises, especially taking into account that these classes respectively represent only 18.96% and 15.49% of the total volume of the PA intensity data available.

Likewise, we acknowledge that data imbalance poses a considerable challenge for the ML learning schemes; whereas at the same time, we believe that the current distribution of data instances per class reflects in a fairly realistic manner what would be to expect in field-use scenarios. Indeed, wearing the sensors in a constant pervasive manner ('quantified-self' paradigm) is very likely to produce situations in which data for low PA intensity are markedly predominant in terms of volume. In fact, it is Exp. A –where users recorded self-selected daily activities freely– which shows the most pronounced imbalance among classes (Table 1).

Regarding the definition of these classes, sports medicine's guidelines for exercise prescription tend to distinguish PA intensity in either five [64] or three stratification levels [45], with both options enjoying wide acceptance in the field. For practical reasons, here we opted for the 3-level convention (i.e. <3, 3–6, >6 MET), instead of for the 5-level system (<2, 2–3, 3–6, 6–8.8, >8.8

MET). Whereas the latter option would allow for a more detailed discernment of PA intensity ranges, it would also imply a notorious non-linear increase in terms of the mathematical complexity of the ML-based pattern learning problem; as well as extra class imbalance. Taking these issues into account, it could be argued whether the potential reward for having information about the two extra ranges (2–3 and 6–8.8 MET) may or may not be worth in practice.

Our PA intensity and modality classifier employed HR measurements directly, i.e. without individual compensations which accounted for subject-specific maximal and/or resting heart rate values (HR$_{max}$, HR$_{rest}$). Data in this regard were not collected during the experiment. In fact, this was a design choice aimed at easing the future deployment of the system in practice. In this sense, removing the burden of having to conduce individualized procedures to determine HR$_{max}$, HR$_{rest}$ with sufficient accuracy, reliability and repeatability (hence getting rid of physiologically normal intra-individual variations) may constitute a remarkable advantage for the monitoring of large-scale populations [9]. Not in vain, the classical rule-of-thumb estimation HR$_{max}$=220-age is widely known to produce considerable errors on an individual basis [12], and reliable determinations of HR$_{max}$ demand maximal stress exercise tests. Whereas accurate determinations of HR$_{max}$, HR$_{rest}$ are an obligatory requirement in certain scenarios (such as for the calibration of accurate HR-based EE estimations [65]), we did not consider them indispensable in the particular scenario of PA intensity and modality stratification tackled here. Nevertheless, methods exist in literature to normalize HR measurements in an individualized manner; thus reducing inter-personal differences in terms of HR values, and the effect of those differences in the outcome of PA classifiers [66].

~~Despite~~ Given the complexity of the ML pipeline and the fairly limited number of participants available for this work, overfitting to data might have been a serious issue threatening the generalization capabilities of our algorithm. With this consideration in mind, during the model proposal stage we took preventive measures against overfitting via the use of nested 10×10-fold CV procedures to tune parameters. In this manner, the optimal number of clusters was chosen as the option which yielded maximal target performance score when evaluated on a separate subfold in the nesting.

Furthermore, to reveal the final potential impact of overfitting in the overall system, we conducted a subsequent LOIOCV analysis. LOIOCV allowed us to assess the recognition performance attained by the ML scheme when tested on an individual whose data were not supplied at all during the training stage, repeating both training and test procedures with each of the participants left aside for evaluation one time, in turns. ~~results~~ Results ~~concerning~~ for PA intensity classification ~~performance~~ as obtained via the ~~LOSOCV~~LOIOCV procedure were virtually identical (performance score 84.65%) to those from ~~the~~ SBSCV, used during model selection (score 84.57±0.54%). In addition, the partial ~~LOSOCV~~LOIOCV ~~results~~ test outcomes when leaving aside subject #A5 (who contributed with as many as 81.87 hours of worth data) ~~out of the training set~~ were highly satisfactory, with a performance score of 96.69%; even though his data were kept out of the training set. Besides, Figure 2 depicts four examples of sequences recognized satisfactorily in a LOIOCV manner. ~~Consequently~~In summary, these ~~two aspects~~ concordant behaviors ~~indicate~~ point towards good generalization capabilities by our algorithm, hence indicating that severe overfitting to the available data is unlikely to have occurred. Nonetheless, this aspect should be confirmed in future validation experiments with larger populations. ~~Hence, overfitting to the available data is unlikely to have occurred.~~

~~Besides, our PA intensity and modality classifier employed HR measurements without the need for subject-specific HR compensations. This point may constitute a remarkable advantage for the monitoring of large populations, removing the burden of conducing individualized calibration procedures [9].~~

In order to establish fair and meaningful evaluations of the performance obtained by the different types of solutions, results were compared [34]~~[34][34][30]~~ against two methods which can be considered the current de facto industry standards in this application domain [51]~~[51][51][44]~~[62]~~[62][62][55]~~. Both methods suffered serious performance problems, particularly to identify vigorous resistance exercise periods from Exp. B. This behavior should not be surprising considering that this type of PA did not involve pronounced waist movements and the mentioned procedures operated only with ActiGraph's accelerometry data, i.e. not incorporating HR information. For this reason, we also deployed two additional ML classifiers (CART trees and NB) to serve as baseline comparisons. With respective scores 72.27% and 76.79% in ~~LOSOCV~~LOIOCV, both methods outperformed the aforementioned approaches without ML [51]~~[51][51][44]~~[62]~~[62][62][55]~~, although they were yet considerably less accurate than our proposal (score 84.65%).

Works [34]~~[34][34][30]~~[35]~~[35][35][31]~~, in the context of developing MET estimators, also addressed the task of ranking PA intensities in an activity-independent manner. Results reported by study [34]~~[34][34][30]~~ (score 75.78%, accuracy 76.99% and *f*-Measures 84.36%, 75.08% and 66.67%) showed a perceptible, yet moderate improvement with respect to the 'cutpoints'-based method by [51]~~[51][51][44]~~ (score: 70.03%) and with the proposal from [62]~~[62][62][55]~~ (score: 71.52%). Nevertheless, this improvement due to their MLP neural network for MET estimation is more modest than in our case. On the other hand, the different criterion adopted by authors in [35]~~[35][35][31]~~ when defining their four intensity intervals, along with their methodology to report results (focused only on recall/sensitivity), prevented us from establishing direct comparisons in terms of performance. Study [28]~~[28][28][24]~~ also addressed activity-independent PA intensity classification, achieving an accuracy of 94.37%. Of note, their least successful identification occurred for moderate ranges, exactly as in this work. In quantitative terms, their performance is higher than the obtained here; however, authors worked on a more homogeneous set of activities (all in laboratory), and employed three accelerometers instead of one. Nonetheless, their study highlights the relevance of applying ML techniques to detect PA patterns when combining accelerometry and HR.

Furthermore, given that most of the classification errors tended to appear around transients in ground truth, instead of as long-term mismatches (Figure 2), the overall impact and potential loss of quality caused by these transient errors should in practice be limited in realistic ambulatory applications. In this regard, the HMM was capable of successfully exploiting temporal redundancy in the data, when long-term trends were maintained stable. However, it had moderate difficulties to match quick fluctuations in the ground truth from the short-interval training protocol from Exp. B, as those transients were up to some extent 'ignored' by the classifier due to the low-pass filtering effect originated in the HMM-based recognition of state transitions. This is in fact the consequence of an intrinsic and unavoidable trade-off between responsiveness to quick changes, on the one side; and ability to benefit from the information contained in long-term trends, on the other. The ML system learnt to favor mainly the latter, as they are more common in the dataset and tended to further improve classification score outcomes. Anyhow, it ~~is~~ can be reasonably expected that those quick oscillations would be rather infrequent in practical free-living scenarios and lifestyle monitoring, where interval training is rare, and self-selected and self-paced exercise is mainly stable.

Besides, the physiological and metabolic adaptations of the body to rapid transients in PA regimes are also slower than those fluctuating changes themselves. This relates to the fact that the organism is effectively responding with certain delay and low-pass response. For example, HR –which constitutes a direct input for our algorithm– returns to basal values at only moderate paces: up to 42 beats/min decays immediately 2 min after exercise termination [64], alongside exponentially slower approaches to basal HR [68]. Moreover, excess post-exercise oxygen consumptions (EPOC) have an important contribution to the total PA-induced energy balance: EPOC may account for up to 6–15% of EE in the exercise session [69]. At the view of these considerations, we do not perceive transient mismatches by the automated PA classifier as a major threat against its applicability in practise.

Another noteworthy practical consideration refers to the relatively high complexity of the pipeline of ML schemes proposed here. In this regard we may note that almost all of the computationally expensive procedures need to be carried out solely during the training stage –when the algorithms learn–; but not in the deployment of the system when applied for recognition purposes, i.e. once an optimal ML model is already selected and trained. Heavy computations (for example: eigenvalue analyses for LDA characterization, the iterative determination of *k*-means cluster centroids, or the estimation of HMM emission and transition matrices; among others) must be performed for the training of the ML system; although these calculations can be carried out at once, e.g. in a powerful desktop computer. Oppositely, the finally deployed PA classifier does not need to repeat them during the ultimate recognition phase, i.e. its normal use in applied practice. Indeed, once the ML scheme has already been trained, computations are not particularly demanding: applying LDA consists in a projection onto a vector subspace (i.e. a matrix multiplication), whereas *k*-means reduces to selecting the minimum scalar value among a set of *k* distances to the corresponding cluster centroids. The only component that might imply certain computational complexity is HMM decoding for temporal filtering, but straightforward efficient implementations exist [60].

Therefore, we would not consider that algorithmic complexity of the ML pipeline or its computational demands –once the system has already been fully trained– should hinder the deployment of the PA classifier in practice. This statement should also hold true even for the case of portable devices with constrained computational power –e.g. a smartphone, with which preliminary pilot tests have already been carried out successfully in our lab–.

Finally, future work should include conducing further validation studies with larger sets of PA data collected from wider populations. This would allow to better characterize the recognition capabilities of our ML solution, to validate the reproducibility of its achievements in terms of performance; as well as to corroborate its hypothesized superiority with respect to other simpler approaches, namely ‘counts’-based thresholds or MLP neural networks alone.

~~7.~~6. Conclusions

This work achieved a robust automatic recognition of PA intensity and exercise modality in an activity-independent manner, by the application of a pipeline of ML techniques to time series of multi-source data incorporating both mechanical (through accelerometry) and physiological (via HR) aspects of exercise.

The automated pattern identification modules and the HMM temporal filtering allowed to gain benefit from the exploitation of time redundancies; reaching classification rates which markedly outperformed standard ‘cutpoints’-based approaches, as well as satisfactory generalization capabilities.

Acknowledgements

The authors thank volunteers for their participation, and members of the PRONAF Study Group for their assistance with data collection.

This work was partly funded by a doctoral research grant from the Universidad Politécnica de Madrid and by the Spanish Ministry of Science and Innovation: research projects ‘APRIORI’ (FIS PS09/01318) and ‘PRONAF’ (DEP-2008-06354-C04-01).

Conflict of interest

No conflict of interest exists.

References

[1] Fogelholm M. Physical activity, fitness and fatness: relations to mortality, morbidity and disease risk factors. A systematic review. Obes Rev 2010; 11(3): 202–221. doi:10.1111/j.1467-789x.2009.00653.x.

[2] Colberg SR, Sigal RJ, Fernhall B, Regensteiner JG, Blissmer BJ, Rubin RR et al. Exercise and type 2 diabetes: The American College of Sports Medicine and the American Diabetes Association, Joint position statement. Diabetes Care 2010; 33(12): e147–e167. doi:10.2337/dc10-9990.

[3] Thompson PD, Buchner D, Piña IL, Balady GJ, Williams MA, Marcus BH et al. Exercise and physical activity in the prevention and treatment of atherosclerotic cardiovascular disease. Circulation 2003; 107(24): 3109–3116. doi:10.1161/01.cir.0000075572.40158.77.

[4] Chavannes N, Vollenberg JJH, van Schayck CP, Wouters EFM. Effects of physical activity in mild to moderate COPD: A systematic review. Br J Gen Pract 2002; 52(480): 574–578.

[5] Holmes MD. Physical activity and survival after breast cancer diagnosis. JAMA 2005; 293(20): 2479–2486. doi:10.1001/jama.293.20.2479.

[6] García-García F, Kumareswaran K, Hovorka R, Hernando ME. Quantifying the acute changes in glucose with exercise in type 1 diabetes: A systematic review and meta-analysis. Sports Med 2015; 45(4): 587–599. doi:10.1007/s40279-015-0302-2.

[7] Breton MD. Physical activity – The major unaccounted impediment to closed loop control. J Diabetes Sci Technol 2008; 2(1): 169–174. doi:10.1177/193229680800200127.

[8] van Bon AC, Verbitskiy E, von Basum G, Hoekstra JBL, de Vries JH. Exercise in closed-loop control: A major hurdle. J Diabetes Sci Technol 2011; 5(6): 1337–1341. doi:10.1177/193229681100500604.

[9] Valanou EM, Bamia C, Trichopoulou A. Methodology of physical activity and energy-expenditure assessment: A review. J Public Health 2006; 14(2): 58–65. doi:10.1007/s10389-006-0021-0.

[10] Freedson PS, Miller K. Objective monitoring of physical activity using motion sensors and heart rate. Res Q Exerc Sport 2000; 71(Suppl2): S21–S29.

[11] Gotshalk LA, Berger RA, Kraemer WJ. Cardiovascular responses to a high-volume continuous circuit resistance training protocol. J Strength Cond Res 2004; 18(4): 760–764.

[12] Ainslie PN, Reilly T, Westerterp KR. Estimating human energy expenditure: A review of techniques with particular reference to doubly labelled water. Sports Med 2003; 33(9): 683–698. doi:10.2165/00007256-200333090-00004.

[13] Plasqui G, Westerterp KR. Accelerometry and heart rate as a measure of physical fitness. Med Sci Sports Exerc 2006; 38(8): 1510–1514. doi:10.1249/01.mss.0000228942.55152.84.

[14] Bassett DR, Rowlands A, Trost SG. Calibration and validation of wearable monitors. Med Sci Sports Exerc 2012; 44(Suppl): S32–S38. doi:10.1249/mss.0b013e3182399cf7.

[15] Bächlin M, Plotnik M, Roggen D, Giladi N, Hausdorff JM, Tröster G. A wearable system to assist walking of Parkinson's disease patients. Benefits and challenges of context-triggered acoustic cueing. Methods Inf Med 2010; 49(1): 88–95. doi:10.3414/ME09-02-0003.

[16] Gietzelt M, Wolf KH, Kohlmann M, Marschollek M, Haux R. Measurement of accelerometry-based gait parameters in people with and without dementia in the field. Methods Inf Med 2013; 52(4): 319–325. doi:10.3414/ME12-02-0009.

[17] Marschollek M, Rehwald A, Wolf KH, Gietzelt M et al. Sensor-based fall risk assessment – An expert 'to go'. Methods Inf Med 2011; 50(5): 420–426. doi:10.3414/ME10-01-0040.

~~[15]~~[18] Bao L, Intille SS. Activity recognition from user-annotated acceleration data. Proceedings of the 2nd International Conference PERVASIVE 2004; 1–17. doi:10.1007/978-3-540-24646-6_1.

[19] Dias A, Gorzelniak L, Schultz K, Wittmann M, Rudnik J et al. Classification of exacerbation episodes in chronic obstructive pulmonary disease patients. Methods Inf Med 2014; 53(2): 108–114. doi:10.3414/ME12-01-0108.

~~[16]~~[20] Ermes M, Parkka J, Mantyjarvi J, Korhonen I. Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions. IEEE Trans Inform Technol Biomed 2008; 12(1): 20–26. doi:10.1109/titb.2007.899496.

~~[17]~~[21] de Vries SI, Garre FG, Engbers LH, Hildebrandt VH, van Buuren S. Evaluation of neural networks to identify types of activity using accelerometers. Med Sci Sports Exerc 2011; 43(1): 101–107. doi:10.1249/mss.0b013e3181e5797d.

~~[18]~~[22] Lau HY, Tong KY, Zhu H. Support Vector Machine for classification of walking conditions using miniature kinematic sensors. Med Biol Eng Comput 2008; 46(6): 563–573. doi:10.1007/s11517-008-0327-x.

~~[19]~~[23] Gyllensten IC, Bonomi AG. Identifying types of physical activity with a single accelerometer: Evaluating laboratory-trained algorithms in daily life. IEEE Trans Biomed Eng 2011; 58(9): 2656–2663. doi:10.1109/tbme.2011.2160723.

~~[20]~~[24] Casale P, Pujol O, Radeva P. Human activity recognition from accelerometer data using a wearable device. Proceedings of the 5th Iberian Conference IbPRIA 2011; 289–296. doi:10.1007/978-3-642-21257-4_36.

~~[21]~~[25] Lester J, Choudhury T, Borriello G. A practical approach to recognizing physical activities. Proceedings of the 4th International Conference PERVASIVE 2006; 1–16. doi:10.1007/11748625_1.

~~[22]~~[26] Tapia EM, Intille SS, Haskell W, Larson K, Wright J, King A et al. Real-time recognition of physical activities and their intensities using wireless accelerometers and a heart rate monitor. Proceedings of the 11th IEEE International Symposium on Wearable Computers 2007. doi:10.1109/iswc.2007.4373774.

~~[23]~~[27] Li M, Rozgic V, Thatte G, Lee S, Emken A, Annavaram M et al. Multimodal physical activity recognition by fusing temporal and cepstral information. IEEE Trans Neural Syst Rehab Eng 2010; 18(4): 369–380. doi:10.1109/tnsre.2010.2053217.

~~[24]~~[28] Reiss A, Stricker D. Aerobic activity monitoring: towards a long-term approach. Univ Access Inf Soc 2013; 13(1): 101–114. doi:10.1007/s10209-013-0292-5.

~~[25]~~[29] Albinali F, Intille S, Haskell W, Rosenberger M. Using wearable activity type detection to improve physical activity energy expenditure estimation. Proceedings of the 12th International Conference on Ubiquitous Computing 2010. doi:10.1145/1864349.1864396.

~~[26]~~[30] Atallah L, Leong JJH, Lo B, Yang GZ. Energy expenditure prediction using a miniaturized ear-worn sensor. Med Sci Sports Exerc 2011; 43(7): 1369–1377. doi:10.1249/mss.0b013e3182093014.

~~[27]~~[31] Liu S, Gao RX, John D, Staudenmayer J, Freedson PS. SVM-based multi-sensor fusion for free-living physical activity assessment. Proceedings of the 33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society EMBS 2011; 3188–3191. doi:10.1109/iembs.2011.6090868.

~~[28]~~[32] Altini M, Penders J, Vullers R, Amft O. Estimating energy expenditure using body-worn accelerometers: A comparison of methods, sensors number and positioning. IEEE J Biomed Health Inform 2015; 19(1): 219–226. doi:10.1109/jbhi.2014.2313039.

~~[29]~~[33] Lin CW, Yang YTC, Wang JS, Yang YC. A wearable sensor module with a neural-network-based activity classification algorithm for daily energy expenditure estimation. IEEE Trans Inform Technol Biomed 2012; 16(5): 991–998. doi:10.1109/titb.2012.2206602.

~~[30]~~[34] Freedson PS, Lyden K, Kozey-Keadle S, Staudenmayer J. Evaluation of artificial neural network algorithms for predicting METs and activity type from accelerometer data: validation on an independent sample. J Appl Physiol 2011; 111(6): 1804–1812. doi:10.1152/japplphysiol.00309.2011.

~~[31]~~[35] Trost SG, Wong WK, Pfeiffer KA, Zheng Y. Artificial neural networks to predict activity type and energy expenditure in youth. Med Sci Sports Exerc 2012; 44(9): 1801–1809. doi:10.1249/mss.0b013e318258ac11.

~~[32]~~[36] Bassett DR, Ainsworth BE, Swartz AM, Strath SJ, O'Brien WL, King GA. Validity of four motion sensors in measuring moderate intensity physical activity. Med Sci Sports Exerc 2000; 32(Suppl): S471–480. doi:10.1097/00005768-200009001-00006.

~~[33]~~[37] Welk GJ, Schaben JA, Morrow JRJ. Reliability of accelerometry-based activity monitors: a generalizability study. Med Sci Sports Exerc 2004; 36(9): 1637–1645.

~~[34]~~[38] Sasaki JE, John D, Freedson PS. Validation and comparison of ActiGraph activity monitors. J Sci Med Sport 2011; 14(5): 411–416. doi:10.1016/j.jsams.2011.04.003.

~~[35]~~[39] Le Masurier GC, Tudor-Locke C. Comparison of pedometer and accelerometer accuracy under controlled conditions. Med Sci Sports Exerc 2003; 35(5): 867–871. doi:10.1249/01.mss.0000064996.63632.10.

~~[36]~~[40] Chen KY, Bassett DR. The technology of accelerometry-based activity monitors: Current and future. Med Sci Sports Exerc 2005; 37(Suppl): S490–S500. doi:10.1249/01.mss.0000185571.49104.82.

~~[37]~~[41] ActiGraph. ActiLife 5 Analysis Software Suite for ActiGraph devices (User manual–Revision H), 2010.

[42] Zapico AG, Benito PJ, González-Gross M, Peinado AB, Morencos E, Romero B et al. Nutrition and physical activity programs for obesity treatment (PRONAF study): Methodological approach of the project. BMC Public Health 2012; 12(1): 1100. doi:10.1186/1471-2458-12-1100.

[43] Byrne NM, Hill AP, Hunter GR, Weinsier RL, Schutz Y. Metabolic equivalent: one size does not fit all. J Appl Physiol 2005; 99(3): 1112–1119.

[44] Wasserman K, Hansen JE, Sue DY, Whipp BJ, Casaburi R. Principles of exercise testing and interpretation. Lippincott Williams & Wilkins, 2 edition, 1994.

~~[38]~~[45] Pate RR, Pratt M, Blair SN, Haskell WL, Macera CA, Bouchard C et al. Physical activity and public health. A recommendation from the Centers for Disease Control and Prevention and the American College of Sports Medicine. JAMA 1995; 273(5): 402–407.

~~[39]~~

[46] Ainsworth BE, Haskell WL, Whitt MC, Irwin ML, Swartz AM, Strath SJ et al. Compendium of physical activities: an update of activity codes and MET intensities. Med Sci Sports Exerc 2000; 32(Suppl): S498–S516. doi:10.1097/00005768-200009001-00009.

~~[40]~~ Benito PJ, Bermejo LM, Peinado AB, López-Plaza B, Cupeiro R, Szendrei B et al. Change in weight and body composition in obese subjects following a hypocaloric diet plus different training programs or physical activity recommendations. J Appl Physiol 2015; 118(8): 1006–1013. doi:10.1152/japplphysiol.00928.2014.

~~[41]~~[47] ~~Pate RR, Pratt M, Blair SN, Haskell WL, Macera CA, Bouchard C et al. Physical activity and public health. A recommendation from the Centers for Disease Control and Prevention and the American College of Sports Medicine. JAMA 1995; 273(5): 402–407.~~

~~[42]~~[48] Preece SJ, Goulermas JY, Kenney LPJ, Howard D, Meijer K, Crompton R. Activity identification using body-mounted sensors – A review of classification techniques. Physiol Meas 2009; 30(4): R1–R33. doi:10.1088/0967-3334/30/4/r01.

~~[43]~~[49] Duda RO, Hart PE, Stork DG. Pattern Classification. Wiley 2000.

~~[44]~~[50] Zumel N, Mount J. Practical Data Science with R. Manning Publications 2014.

~~[45]~~[51] Freedson PS, Melanson E, Sirard J. Calibration of the Computer Science and Applications, Inc. accelerometer. Med Sci Sports Exerc 1998; 30(5): 777–781. doi:10.1097/00005768-199805000-00021.

~~[46]~~[52] Rencher AC, Christensen WF. Methods of Multivariate Analysis. John Wiley & Sons 2012.

~~[47]~~[53] Kohavi R, John GH. Wrappers for feature subset selection. Artificial Intelligence 1997; 97(1): 273–324. doi:10.1016/s0004-3702(97)00043-x.

~~[48]~~[54] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003; 3: 1157–1182.
~~[49]~~[55] Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Machine Intell 2005; 27(8): 1226–1238. doi:10.1109/tpami.2005.159.
~~[50]~~[56] Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. J Bioinform Comput Biol 2005; 03(02): 185–205. doi:10.1142/s0219720005001004.
~~[51]~~[57] Pober DM, Staudenmayer J, Raphael C, Freedson PS. Development of novel techniques to classify physical activity mode using accelerometers. Med Sci Sports Exerc 2006; 38(9): 1626–1634. doi:10.1249/01.mss.0000227542.43669.45.
~~[52]~~[58] He J, Li H, Tan J. Real-time daily activity classification with wireless sensor networks using Hidden Markov Model. Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society EMBS 2007. doi:10.1109/iembs.2007.4353008.
~~[53]~~[59] Mannini A, Sabatini A. M. Accelerometry-based classification of human activities using Markov modeling. Comput Intell Neurosci 2011; 2011:1–10. doi:10.1155/2011/647858.
~~[54]~~[60] Rabiner L. A tutorial on Hidden Markov Models and selected applications in speech recognition. Proceedings of the IEEE 1989; 77(2): 257–286. doi:10.1109/5.18626.
~~[55]~~[61] Forman G, Scholz M. Apples-to-apples in cross-validation studies. ACM SIGKDD Explor Newsl 2010; 12(1): 49. doi:10.1145/1882471.1882479.
[62] Crouter SE, Churilla JR, Bassett DR. Estimating energy expenditure using accelerometers. Eur J Appl Physiol 2006; 98(6): 601–612. doi:10.1007/s00421-006-0307-5.
~~[56]~~
[63] Brage S, Brage N, Franks PW, Ekelund U, Wareham NJ. Reliability and validity of the combined heart rate and movement sensor Actiheart. Eur J Clin Nutr 2005; 59(4): 561–570. doi:10.1038/sj.ejcn.1602118.
[64] American College of Sports Medicine. ACSM's guidelines for exercise testing and prescription. Lippincott Williams & Wilkins, 2013.
[65] Wareham NJ, Rennie KL. The assessment of physical activity in individuals and populations: why try to be more precise about how physical activity is assessed? Int J Obes Relat Metab Disord 1998; 22(Suppl 2): S30–8.
[66] Altini M, Penders J, Vullers R, Amft O. Automatic heart rate normalization for accurate energy expenditure estimation. Methods Inf Med 2014; 53(5): 382–8. doi:10.3414/ME13-02-0031.
[67] Froelicher VF, Myers J. Exercise and the heart. 5th ed. Saunders 2006.
[68] Coote JH. Recovery of heart rate following intense dynamic exercise. Exp Physiol 2010; 95(3): 431–440. doi: 10.1113/expphysiol.2009.047548.
~~[57]~~[69] Hall KD, Heymsfield SB, Kemnitz JW, Klein S, Schoeller DA, Speakman JR. Energy balance and its components: implications for body weight regulation. Am J Clin Nutr 2012; 95(4): 989–94. doi:10.3945/ajcn.112.036350.

Figures

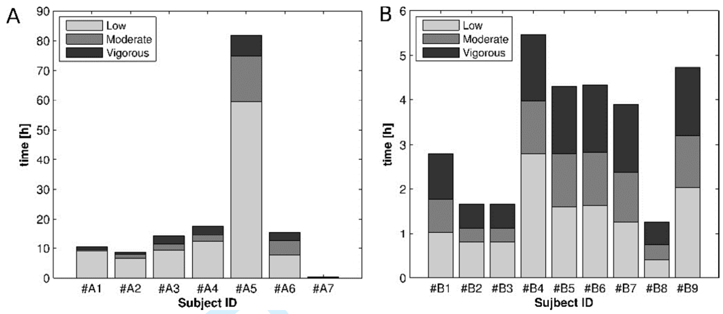

Figure 1. Individual contributions to the dataset about PA intensity; divided by experiments (panels A, B), by participants and by intensity levels.

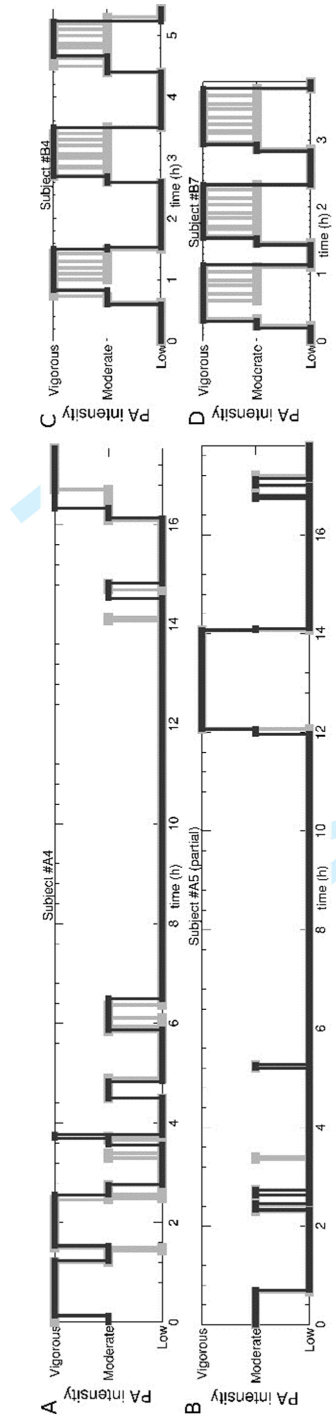

Figure 2. Four example outcome sequences, automatically generated during ~~LOSOCV~~LOIOCV classification by the LDA+*k*-means+HMM scheme (dark line); plotted versus the evolution of ground truth (light line) over time. Panels A, B (left) correspond to free-living activity (Exp. A); whereas on the contrary, panels C, D (right) depict data captured during Exp. B.

Formatted: Font: Italic

Tables

Table 1. Summary of the dataset concerning PA intensity.

~~PA intensity~~	Experiment A 7 participants 72 sessions 148.50 h (83.13%)	Experiment B 9 participants 20 sessions 30.13 h (16.87%)	Total 16 participants 92 sessions 178.63 h (100.00%)
Low (<3 MET)	104.73 h (70.53%)	12.37 h (41.04%)	117.10 h (65.55%)
Moderate (3–6 MET)	26.30 h (17.71%)	7.57 h (25.11%)	33.87 h (18.96%)
Vigorous (>6 MET)	17.47 h (11.76%)	10.20 h (33.85%)	27.67 h (15.49%)

Table 2. Model selection – SBSCV performances in the task of PA intensity classification.

mean±SD (n=30)		Dimensionality reduction: PCA (8 features)				Dimensionality reduction: LDA (2 features)			
		***k*-means** (45 clusters)	**GMM** (36 clusters)	**Hierarchical** (28 clusters)	**SOM** (43 clusters)	***k*-means** (26 clusters)	**GMM** (25 clusters)	**Hierarchical** (18 clusters)	**SOM** (24 clusters)
Accuracy		84.67±0.64%	86.85±0.56%	84.59±0.63%	84.84±0.64%	88.83±0.45%	88.23±0.60%	88.81±0.45%	89.09±0.42%
***f*-Measure**	Low	93.54±0.51%	94.50±0.31%	93.38±0.44%	93.71±0.42%	95.56±0.31%	95.37±0.36%	95.58±0.31%	95.85±0.25%
	Moderate	64.61±1.15%	68.35±1.15%	66.19±1.46%	65.90±1.52%	71.90±1.11%	70.50±1.46%	71.95±1.05%	72.22±1.16%
	Vigorous	74.34±1.49%	78.76±1.55%	72.53±1.92%	72.50±1.47%	82.01±0.65%	80.76±1.11%	81.77±0.84%	81.38±0.77%
Score		79.29±0.72%	82.11±0.76%	79.17±0.76%	79.24±0.82%	84.57±0.54%	83.72±0.78%	84.53±0.57%	84.67±0.57%

mean±SD (n=30)		Dimensionality reduction: mRMR continuous (8 features)				Dimensionality reduction: mRMR discrete (8 features)			
		***k*-means** (37 clusters)	**GMM** (45 clusters)	**Hierarchical** (48 clusters)	**SOM** (45 clusters)	***k*-means** (31 clusters)	**GMM** (43 clusters)	**Hierarchical** (35 clusters)	**SOM** (42 clusters)
Accuracy		87.09±0.61%	87.24±0.61%	87.36±0.59%	87.47±0.43%	86.75±0.35%	86.48±0.47%	86.66±0.54%	87.57±0.57%
***f*-Measure**	Low	95.17±0.37%	95.15±0.30%	95.41±0.39%	95.43±0.33%	95.43±0.27%	94.43±0.28%	95.06±0.31%	95.56±0.40%
	Moderate	67.86±1.26%	67.76±1.42%	68.10±1.36%	68.49±1.03%	66.64±0.91%	68.19±1.29%	66.84±1.20%	68.71±1.21%
	Vigorous	77.82±1.20%	78.64±1.31%	77.85±1.01%	77.80±1.00%	75.97±0.88%	76.56±1.29%	76.96±1.06%	76.91±1.10%
Score		81.99±0.76%	82.20±0.85%	82.18±0.72%	82.30±0.55%	81.20±0.44%	81.41±0.67%	81.38±0.68%	82.19±0.68%

Table 3. SBSCV performance by the LDA+*k*-means+HMM algorithm.

mean±SD (n=30)	**PA intensity** (10-fold SBSCV)			**PA modality** (5-fold SBSCV)		
	Low	Moderate	Vigorous	Sustained aerobic	Mixed	Resistance
Accuracy		88.83±0.45%			83.37±6.93%	
Precision	97.23±0.18%	72.32±1.47%	76.44±0.77%	81.33±6.03%	77.22±17.02%	92.35±3.45%
Recall	93.94±0.51%	71.49±1.17%	88.46±0.90%	84.89±16.51%	77.94±7.33%	84.70±5.69%
***f*-Measure**	95.56±0.31%	71.90±1.11%	82.01±0.65%	82.36±10.95%	76.46±10.60%	88.26±3.83%
Score		84.57±0.54%			82.61±7.16%	

Table 4. ~~LOSOCV~~LOIOCV confusion matrix for PA intensity classification, as attained by the LDA+*k*-means+HMM combination of algorithms. Each data item corresponds to a 2-min window.

	Classification outcome		
Ground truth	Low	Moderate	Vigorous
Low	3297	194	22
Moderate	83	743	190
Vigorous	5	103	722

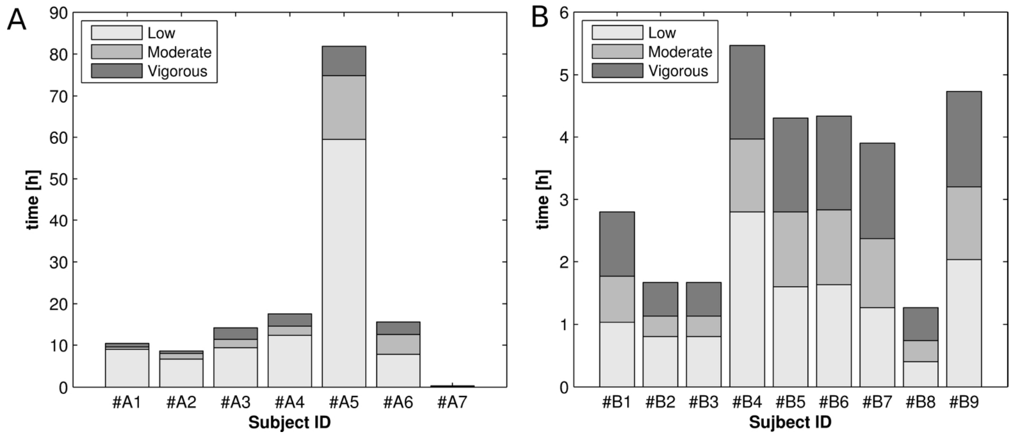

Figure 1. Individual contributions to the dataset about PA intensity; divided by experiments (panels A, B), by participants and by intensity levels.
88x37mm (600 x 600 DPI)

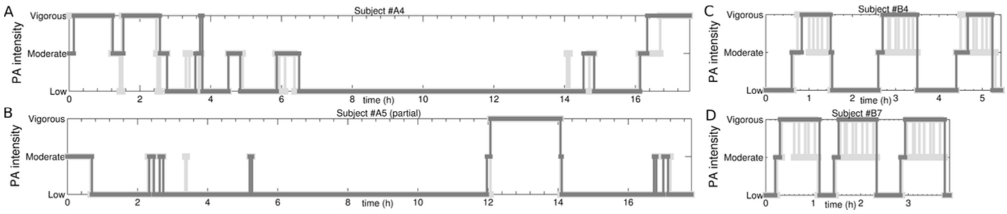

Figure 2. Four example outcome sequences, automatically generated during LOIOCV classification by the LDA+k-means+HMM scheme (dark line); plotted versus the evolution of ground truth (light line) over time. Panels A, B (left) correspond to free-living activity (Exp. A); whereas on the contrary, panels C, D (right) depict data captured during Exp. B.
52x10mm (600 x 600 DPI)